Two type of projects:

1. Scaling of model training runtime and test accuracy with increasing dataset size
    a. Pick model
    b. Pick two large datasets
    c. Use random sampling to generate datasets of different sizes
    d. Perform cross validation across each dataset size

2. Comparison of two models: test accuracy and training runtime
    a. Pick two models
    b. Pick two large datasets
    c. Perform cross validation

Linear models: Linear regression and support vector machine

Non-linear models: Neural networks

Datasets:

1. UCI machine learning repository: https://archive.ics.uci.edu/
2. Kaggle: https://www.kaggle.com/
3. Google Datasets: https://datasetsearch.research.google.com/
4. Papers with code: https://paperswithcode.com/

Problem domains:

1. Image classification - popular benchmarks are CIFAR10, CIFAR100, IMAGENET, STL10, MNIST
2. Video classification
3. Image segmentation
4. Image localization
5. Tabular data - business data such as insurance, time series data